Emerging Database Technologies

Duration : 04hrs

Distributed DBMS

Distributed DBMS

- Centralised DBMS all system components (data, DBMS software and secondary storage devices) reside at a single computer or site. A centralised database can be accessed remotely via terminals connected to the site.

- In Distributed DBMS, each site is a database system site in its own right, but the sites have agreed to work together (if necessary). User at any site can access data anywhere in the network exactly as if the data were all stored at user's own site.

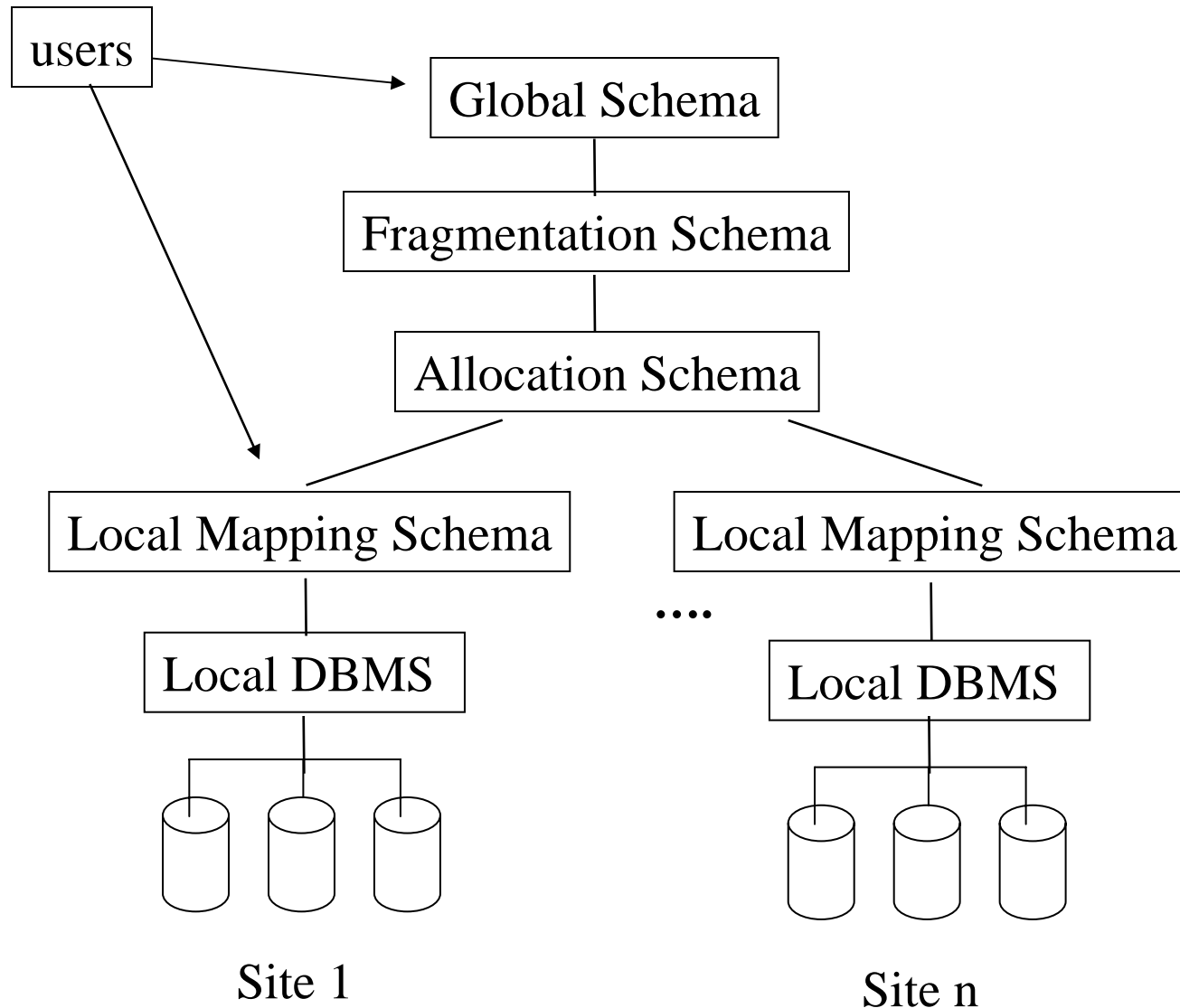# Motivation for Distributed Databases

- Organisational and economic reason - Many organisations are decentralised and a distributed database approach fits more naturally the structure of the organisation. E.g. Banks.

- Interconnection of existing databases - Distributed databases are a natural solution when several databases already exist in an organisation and the necessity of performing global applications arises.

- Incremental growth - Supports organisational growth (new branches) in a smoother manner than with a centralised database approach.

- Allow data sharing while maintaining some measure of local control (autonomy).

Motivation for Distributed Databases

- Reduce Communication overhead - This is not automatically guaranteed by distribution, but depends largely on the quality of the distributed database design.

- Performance Consideration - The existence of several processors results in better performance through the use of parallelism. Smaller databases exist at each site and hence, local queries and transactions are improved.

- Increased reliability and availability - The use of multiple components means that higher reliability can be obtained. Also data replication can be used to increase availability of data.

General Architecture for a Distributed Databases

BIT

users → Global Schema

Global Schema
|
Fragmentation Schema
|
Allocation Schema
|
Local Mapping Schema   Local Mapping Schema
| .... |
Local DBMS   Local DBMS

Site 1   Site n

Global Schema Defines all the data contained in the distributed database as if the data were not distributed at all. It consists of a set of global relations.

UCSC

General Architecture for a Distributed Databases

Fragmentation Schema - Each global relation can be split into several non-overlapping portions called fragments. The mapping between global relations and fragments is defined in the fragmentation schema. Here, several fragments correspond to one global relation.

Allocation Schema - Fragments are physically located at one or several sites of the network. The mapping defined in the allocation schema determines whether the distributed database is redundant or non-redundant. Each allocation corresponds to a fragment if the data is non-redundant otherwise several allocations will correspond to a fragment.

Local Mapping Schemes - The top three levels are site independent. They do not depend on the data model of the local DBMS. At the lower level, it is necessary to map the objects to those, which are manipulated by the local DBMS. This mapping is called the local mapping schema.

Data Warehouse

Online transaction processing (OLTP) systems

Designed to get data in quickly and to analyse current events.

Characterised by:

- Process oriented

- Data Normalised

- Current data

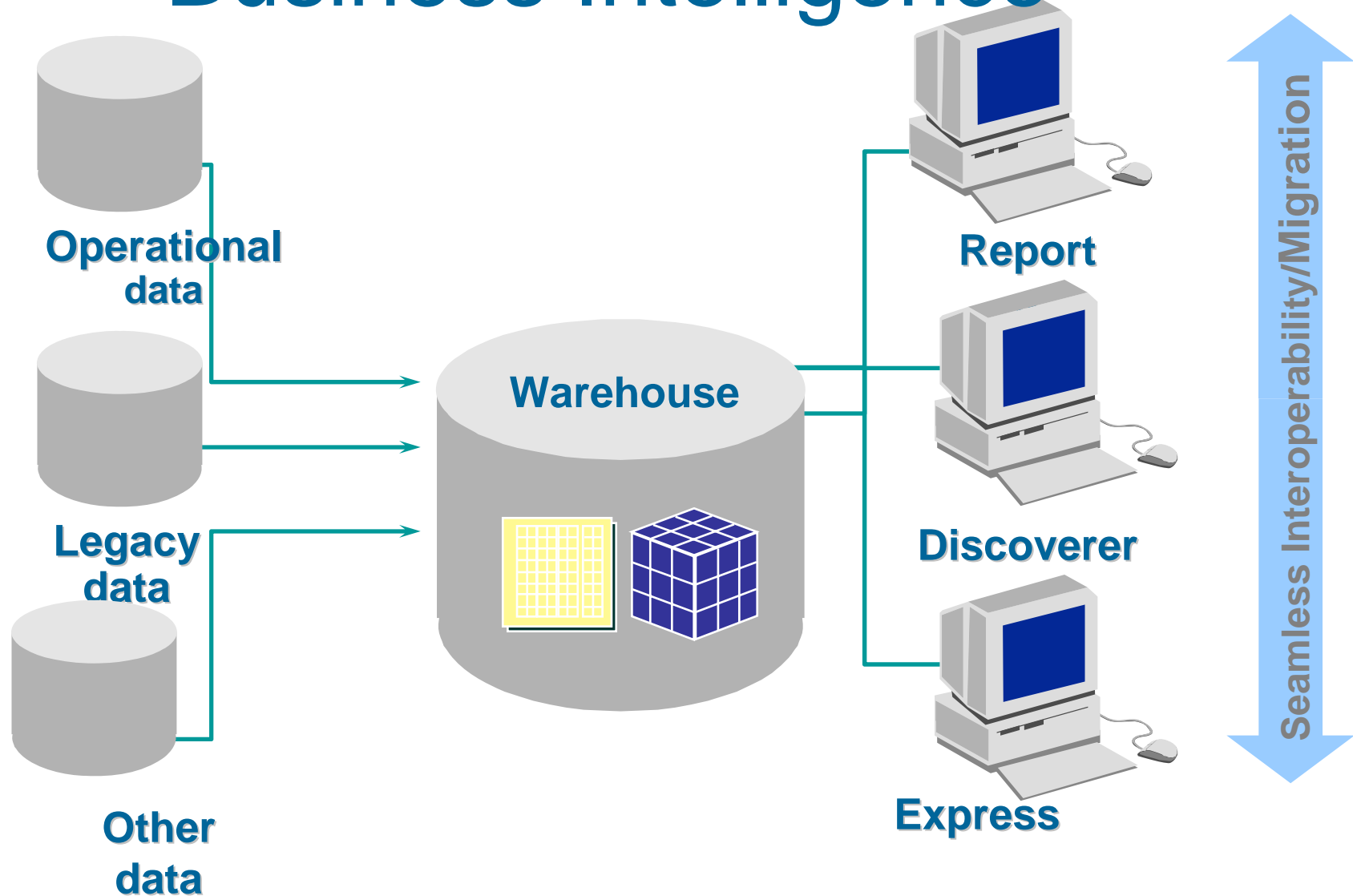- Volatile data

- Updated in real-time
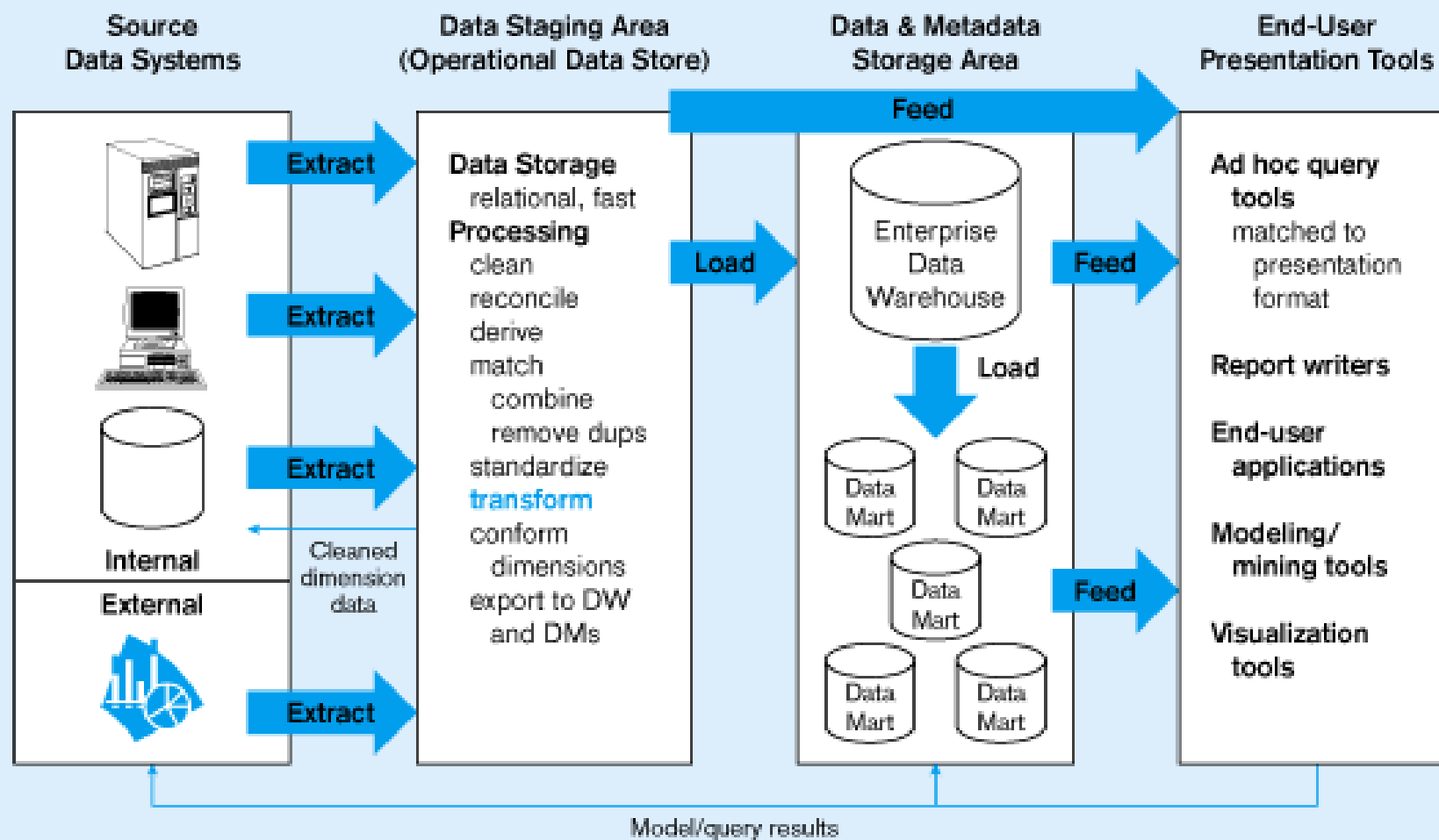
Data Warehouse Systems

Designed to get data out and quickly analyse.

Characterised by:
- Subject oriented rather than process orientated
- Integrated across subjects and entire enterprise
- De-normalised data
- Time-variant
- Historical
- Non Volatile
- Atomic and Summary data

An Integrated Environment for Business Intelligence

Operational data

Legacy data

Other data

Warehouse

Report

Discoverer

Express

Seamless Interoperability/Migration

© 2007, UCSC

© 2007, UCSC

Data Mining

What Is Data Mining?

- Data mining (knowledge discovery in databases):
  - Extraction of interesting (<u>non-trivial,</u> <u>implicit,</u> <u>previously unknown</u> and <u>potentially useful</u>) information or patterns from data in <u>large databases</u>
- What is not data mining?
  - (Deductive) query processing.
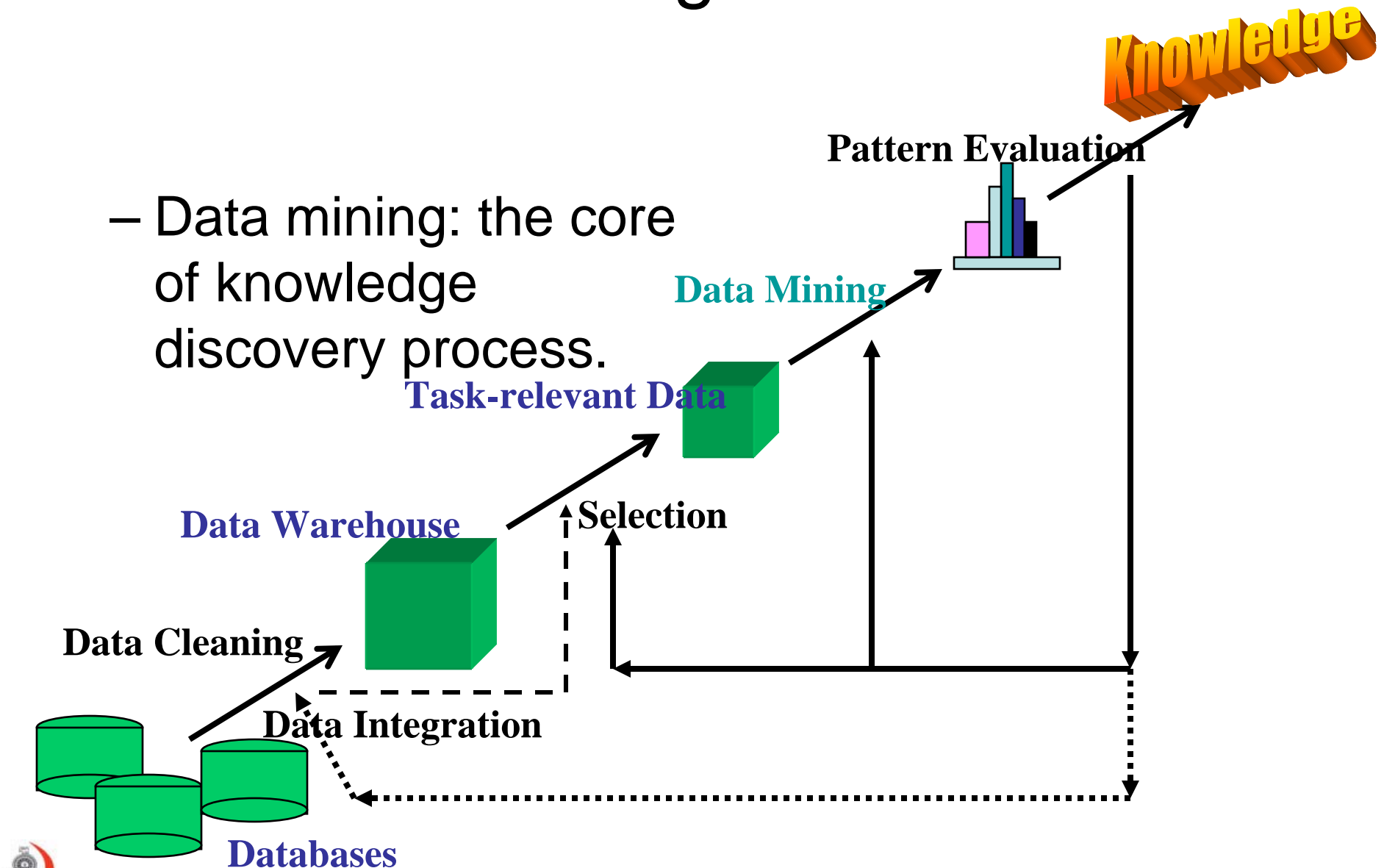 - Expert systems or small ML/statistical programs

Why Data Mining? — Potential Applications

- Database analysis and decision support
  - Market analysis and management
    - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and management

Data Mining: A KDD Process

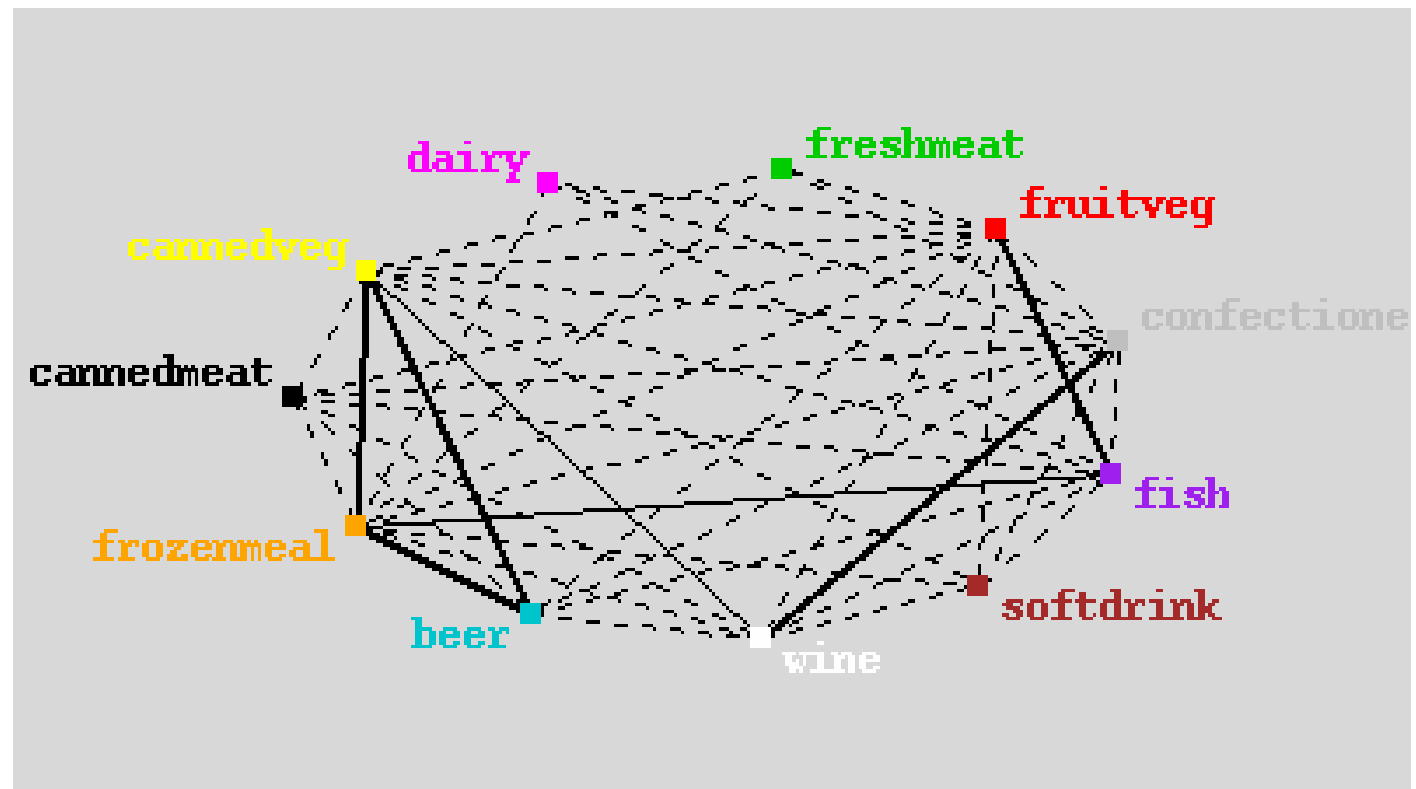– Data mining: the core of knowledge discovery process.

Knowledge

Pattern Evaluation

Data Mining

Task-relevant Data

Data Warehouse

Selection

Data Cleaning

Data Integration

Databases

Association

Description

- Seeks *association rules* in dataset
- 'Market basket' analysis
- Sequence discovery

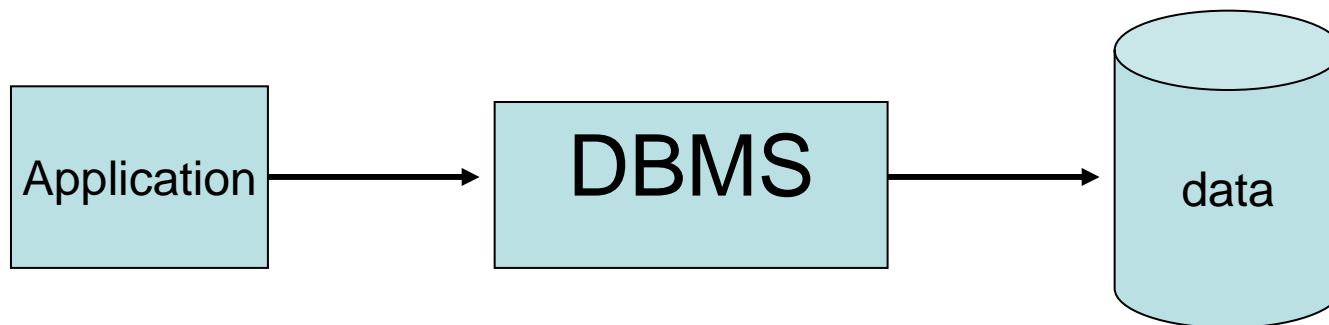- **Open Systems, Interoperability, Database access over Internet, Open Database Connectivity (ODBC)**
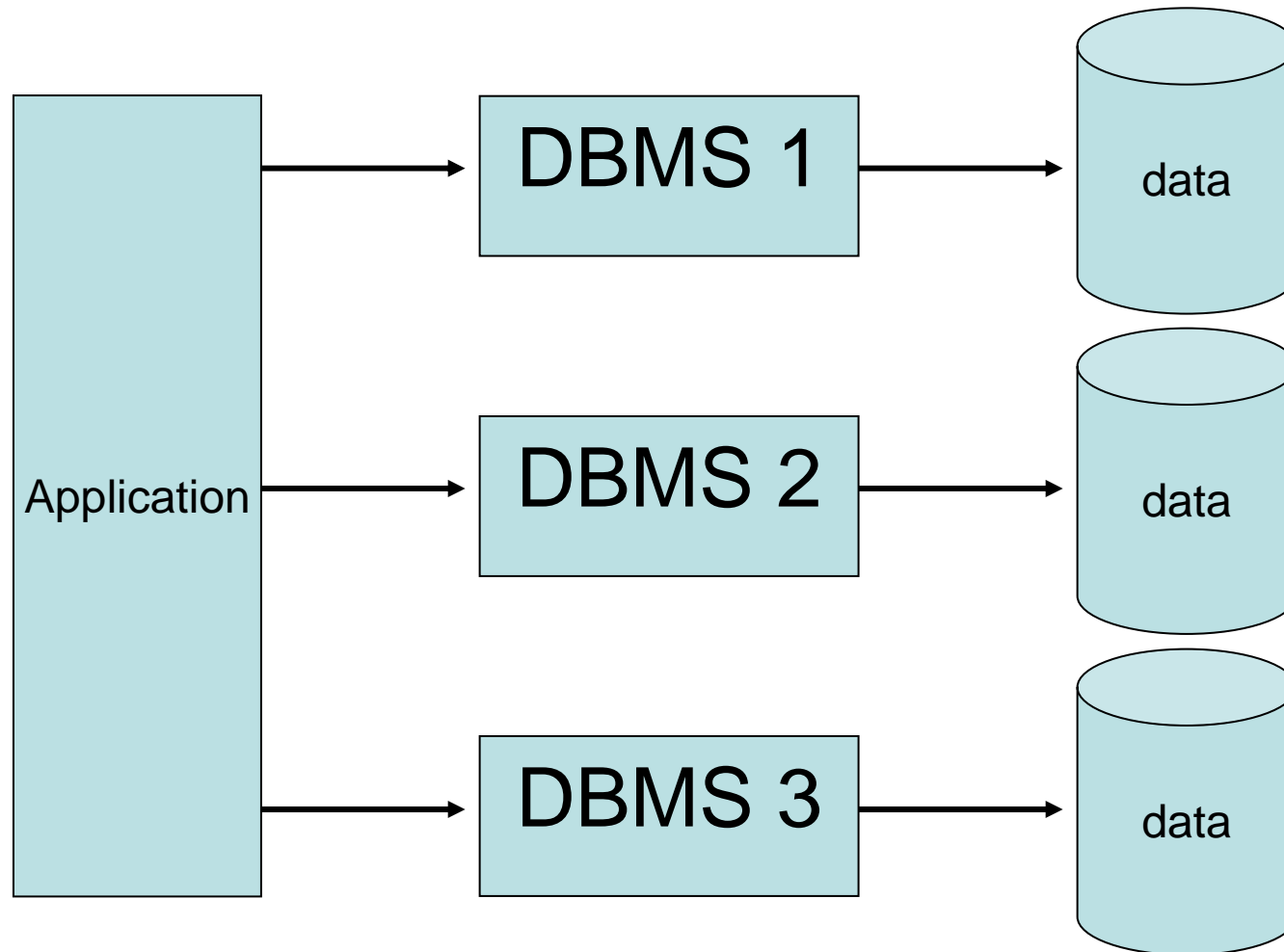
Open Systems

- Standards for Database Interoperability and Portability allows systems to be open.

- Access to different data sources would be achieved through standard SQL based applications.
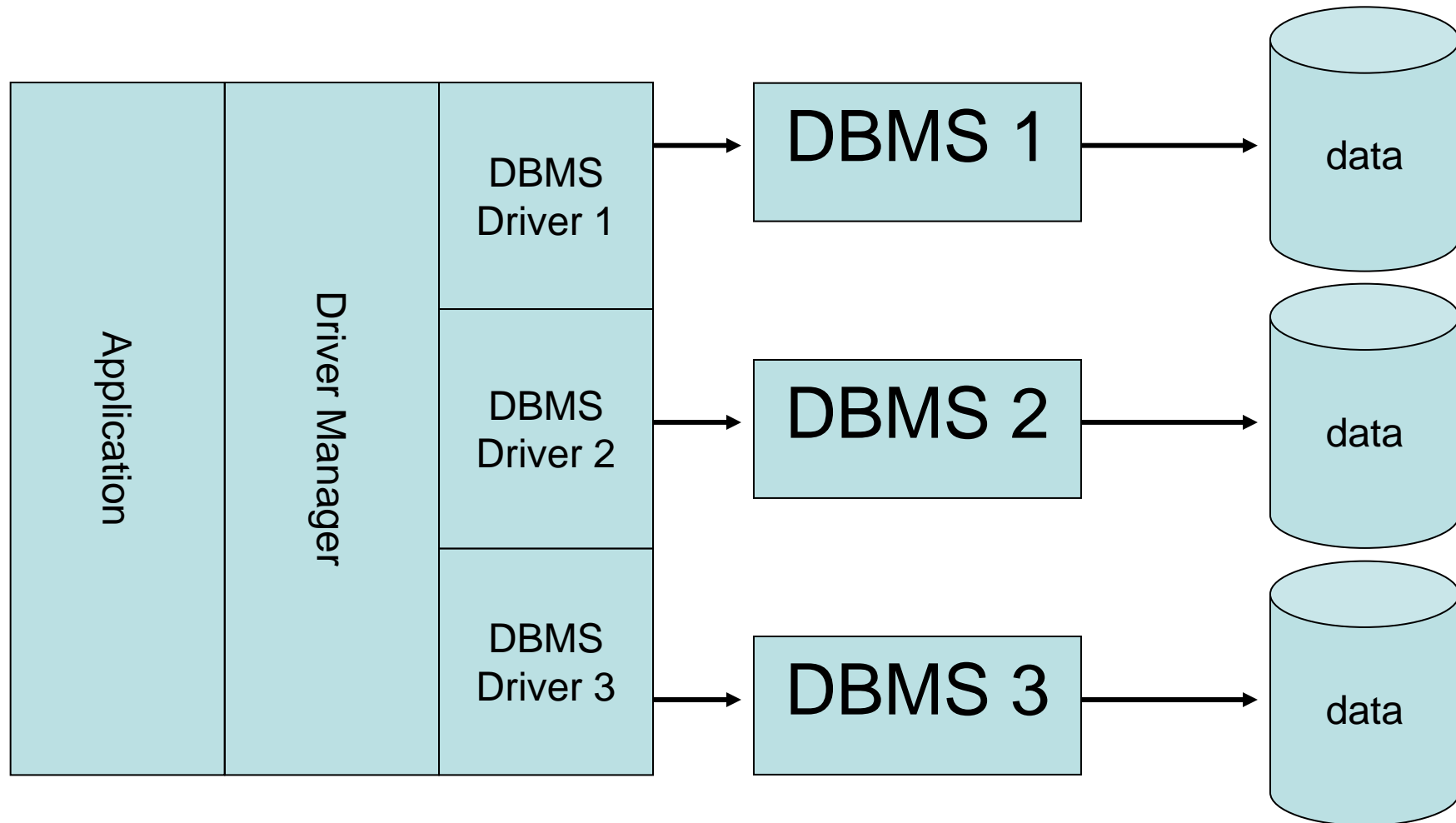
Database Connectivity

```
+-------------+          +-------------+          +---------+
| Application | -------> |    DBMS     | -------> |  data   |
+-------------+          +-------------+          +---------+
```

Simple Database Application

Database Connectivity

Multi-Database Application

Database Connectivity



Standard Access to Databases

© 2007, UCSC

Open Database Connectivity (ODBC) Standard

- ODBC is an interface by which an application programs can access and process SQL databases in a DBMS-independent manner.
- It contains Four components
  - Application (when an application is written in Java connectivity is via JDBC API)
  - Driver Manager (is supplied by the vendor of the OS where an application is running)
  - DBMS Driver (supplied by DBMS vendor or independent software company)
  - Data Source (database associated with OS, network etc)

Discoverer

**"An ad hoc query, reporting, and analysis tool"**
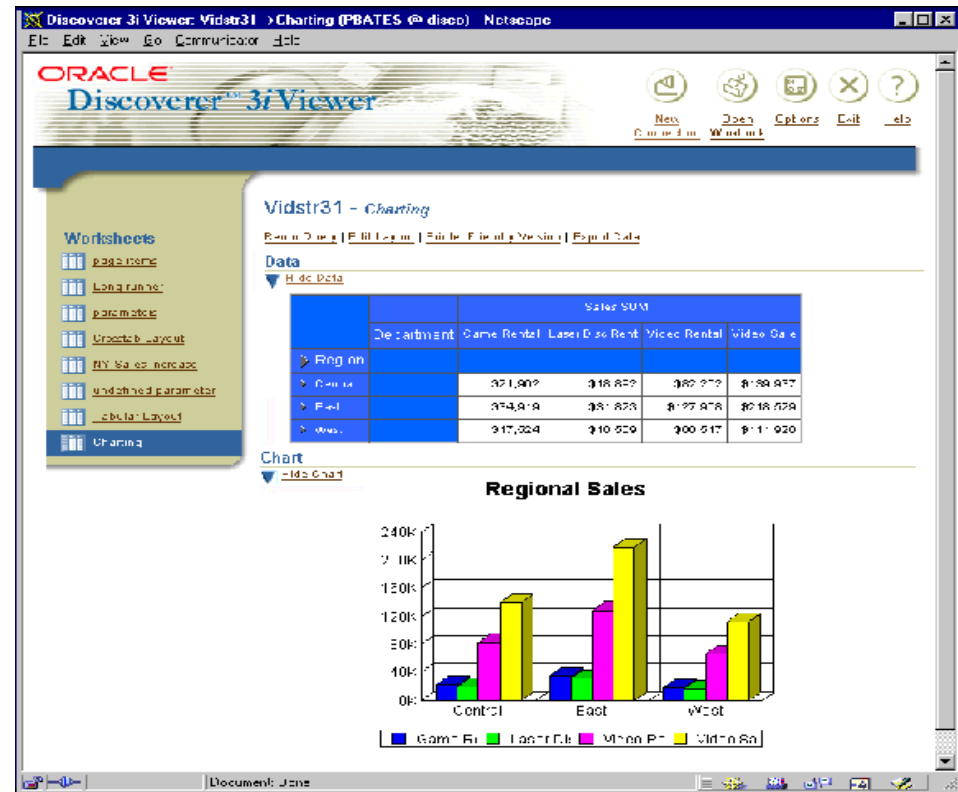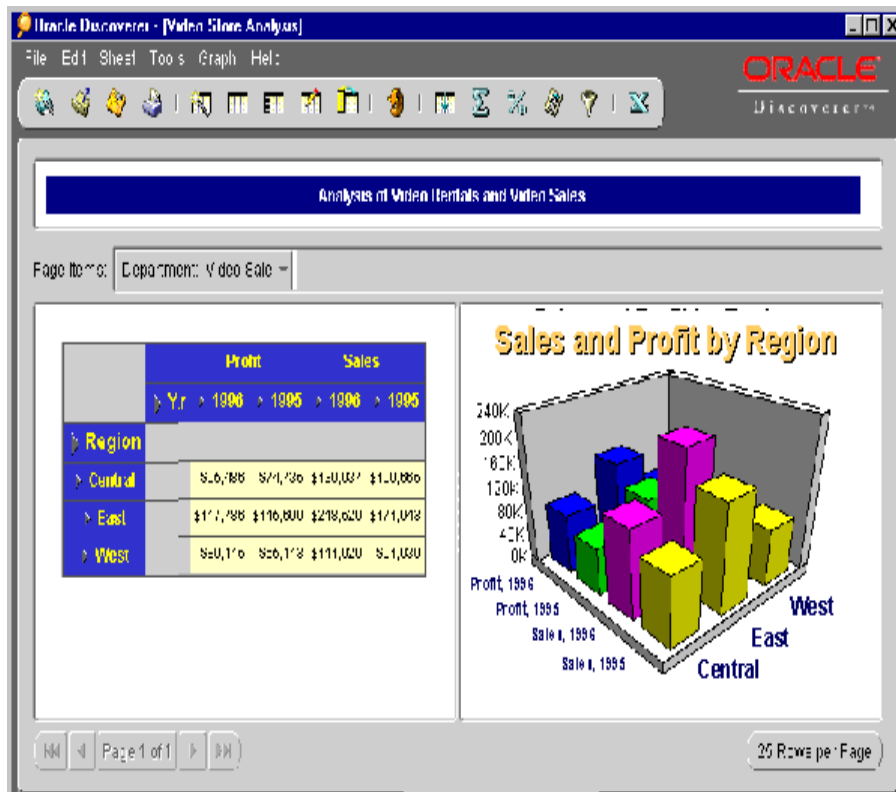
XML Database

- Ultra small database – less than 50 Kb
- Handle large data-volumes
- Search, insert, update, and delete the data in the database created by XML file
- The database can be part of an application
- JDBC API is available

Capabilities

- Search, insert, update, and delete the data in the database created by XML file
- The database can be part of an application
- JDBC API is available

Structure
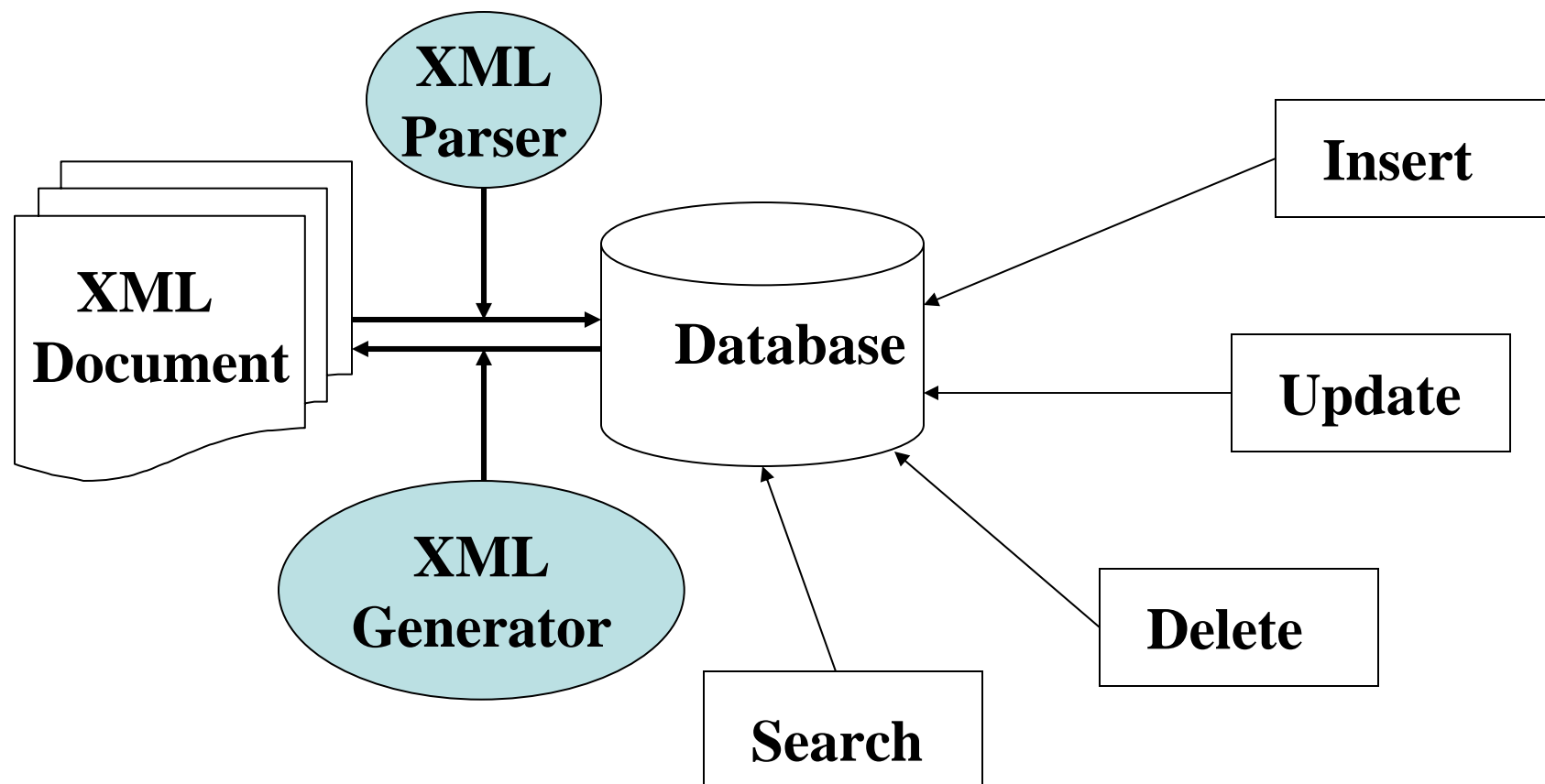
Sample XML File

```
<Telephone>
 <EntryID>1038</EntryID>
 <LoginName>jake</LoginName>
 <PassWord>pass38</PassWord>
 <Lastname>Kim</Lastname>
 <Firstname>Jungkee</Firstname>
 <Date_of_Birth>10.01.1964</Date_of_Birth>
 <Company>FSU</Company>
 <Salutation>Mr.</Salutation>
 <Email>jake@csit.fsu.edu</Email>
 <Address><Street>400 Dirac Science Library</Street>
         <City>Tallahassee</City>
         <ZIP>32306</ZIP>
         <Country>Korea</Country>
         <Telephone>6447018</Telephone>
         <Fax />
   </Address>
 </Telephone>
```